

## Prática 6 - Análise de Dados usando e pandas

1. Análise de Dados
2. Formatos tabulares (CSV e Excel)
3. Utilização das classes Series e DataFrame do módulo pandas em processamento de dados

**Recomendação:** Apesar de este guião poder ser resolvido no seu editor ou IDE habitual, recomenda-se a utilização de um notebook.

### Exercício 6.1 - Atletas

Considerando o ficheiro [Athletes.csv](#), usando Pandas, dê resposta às questões:

1. Quantas colunas tem o ficheiro? Qual a primeira? E a última?
2. Quantas linhas tem o ficheiro?
3. Obtenha um ficheiro de dados em formato Excel com apenas as primeiras 10 colunas. Na resposta às questões seguintes considere apenas essas 10 colunas.
4. Faça com que o índice seja o nome e mostre informação dos primeiros cinco atletas.
5. Altere o tipo da coluna "Sport" para o tipo mais adequado. Repita o processo para todas as colunas em situação similar.
6. Altere os nomes das colunas para as traduções dos originais para Português.
7. O ficheiro tem atletas de quantos desportos? Quais são esses desportos? Qual o desporto com mais atletas no ficheiro?
8. Ordene por país e mostre apenas o nome, desporto e país para os 7 primeiros. Esses 7 praticam que desportos?
9. Que tipos de dados têm as colunas relacionadas com rendimentos? Converta-as para inteiros e determine os 3 atletas com maiores rendimentos.
10. Qual o rendimento médio global? E por género? E por desporto?
11. Crie uma função que permita obter uma nova DataFrame com apenas os atletas com rendimento superior a um valor passado como parâmetro. A função deve devolver as linhas ordenadas por ordem decrescente de "Total Pay" e "Salary/Winnings". Utilize-a para obter tabelas para rendimentos superiores as 10 milhões, 30 e 70 milhões de dólares. Quantos atletas encontrou no último caso?
12. Crie uma nova questão que possa ser respondida com os dados no ficheiro e obtenha a resposta.

## Exercício 6.2 - Proveitos de Turismo em Portugal

Considerando a informação relativa aos proveitos de alojamentos turísticos em Portugal de 2009 a 2020 disponibilizada pela PORDATA (Folha “Quadro” do ficheiro [pordata1.xls<sup>4</sup>](#)), responda às seguintes questões usando as funcionalidades da **biblioteca Pandas**:

1. Leia os dados do ficheiro para um objeto DataFrame com o nome `pordata`, descartando a primeira linha e fazendo com que o índice seja a Região.
2. Determine o número de colunas, número de linhas e memória ocupada pela DataFrame sem otimizações (em kB).
3. Crie uma nova DataFrame com apenas a informação relativa às NUTS II e total de proveitos para os vários anos. Guarde-a num ficheiro CSV.
4. Qual o valor total e valor médio do total de proveitos em 2009 e 2020 nas regiões NUTS II?
5. Obtenha as estatísticas principais para as regiões NUTS II em 20230 para o total e apenas considerando os hotéis.
6. Obtenha o valor médio, máximo, mínimo e soma para os municípios de Aveiro, Porto, Albufeira e Funchal para os proveitos dos hoteis entre 2009 e 2020. Apresente os resultados ordenados por ordem crescente do valor médio, uma cidade por linha.
7. (opcional) Utilize a conversão de tipos de dados para diminuir a memória ocupada. Quantos kBytes conseguiu poupar?

## Exercício 6.3- Netflix

O ficheiro de dados [netflix.csv](#) contém informação sobre cerca de 6000 títulos disponíveis para visualização no serviço de streaming da Netflix em Novembro de 2019. O ficheiro tem 4 colunas: título do vídeo, realizador, data em que foi adicionado ao catálogo da empresa e tipo/categoria. As colunas realizador e data têm alguns valores em falta. As primeira 5 linhas são:

	title	director	date_added	type
0	Alias Grace	NaN	2017-11-03	TV Show
1	A Patch of Fog	Michael Lennox	2017-04-15	Movie
2	Lunatics	NaN	2019-04-19	TV Show
3	Uriyadi 2	Vijay Kumar	2019-08-02	Movie
4	Shrek the Musical	Jason Moore	2013-12-29	Movie

Resolva os seguintes desafios:

1. Leia os dados para uma DataFrame pandas e optimize a utilização de memória e utilidade dos dados pelo processamento das datas e colunas com um número limitado de valores (categorias).
2. Altere a designação das colunas para português (título, realizador, data\_adicionado, tipo).

---

<sup>4</sup> O ficheiro inclui colunas para os anos de 2009 a 2020 dos proveitos totais e apenas para Hotéis. A coluna Região ....

Confirme que a alteração teve sucesso visualizando as primeiras 3 linhas dos dados

3. Quantos nomes de filmes diferentes contêm os dados? Encontre todas as linhas com o título "Supergirl".
4. Quantos realizadores de programas de televisão diferentes contêm os dados? Quais os realizadores de TV que também têm trabalhos em cinema? Encontre todas as linhas do realizador "Robert Rodriguez" e tipo "Movie".
5. Encontre todas as linhas adicionadas em "2019-07-31" ou realizador igual a "Robert Altman". E se, mantendo o realizador, estivermos interessados em adicionados nos anos de 2018 e 2019?
6. Encontre todas as linhas dos realizadores "Orson Welles", "Steven Spielberg", "Sam Raimi" ou "Onir" e mostre uma lista com os filmes.
7. Encontre todas as linhas adicionadas entre 1 de maio de 2019 e 1 de junho de 2019, excluindo esta última data.
8. Remova todas as linhas com um valor NaN na coluna do realizador. Quantos registos passou a ter a sua DataFrame?
9. Identifique os dias em que a Netflix adicionou apenas um filme ao seu catálogo.

#### Exercício 6.4 - NFL

O ficheiro [nfl.csv](#) contém uma lista de jogadores na Liga Nacional de Futebol dos Estados Unidos (NFL), com as colunas: nome, equipa, posição, aniversário e salário. Responda às perguntas seguintes:

1. Como podemos importar o ficheiro .csv NFL? Qual é uma forma eficaz de converter os valores na sua coluna de aniversário para o tipo data?
2. Quais são as duas formas de definir o índice DataFrame para armazenar os nomes dos jogadores?
3. Como podemos contar o número de jogadores por equipa neste conjunto de dados?
4. Quem são os cinco jogadores mais bem pagos?
5. Como podemos ordenar os dados primeiro por equipas por ordem alfabética e depois por salário em ordem descendente?
6. Quem é o jogador mais velho da lista dos New York Jets, e qual o dia do seu aniversário?